Introduction to Machine Learning
Lecture 8

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Support Vector Machines

This Lecture

- Support Vector Machines - separable case
- Support Vector Machines - non-separable case
- Margin guarantees

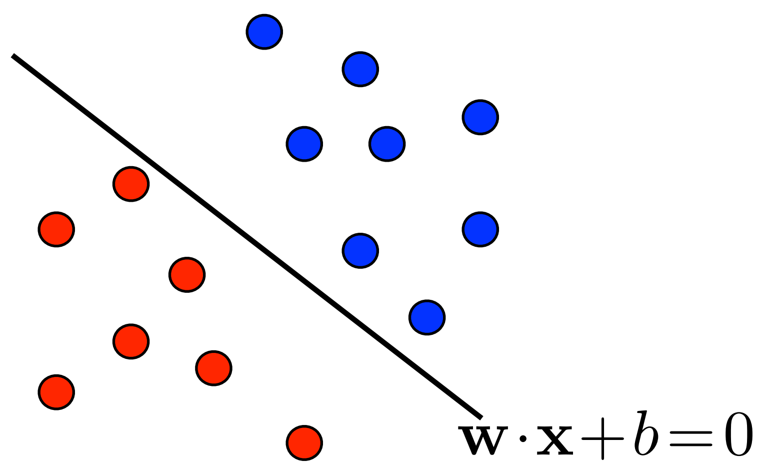
Binary Classification Problem

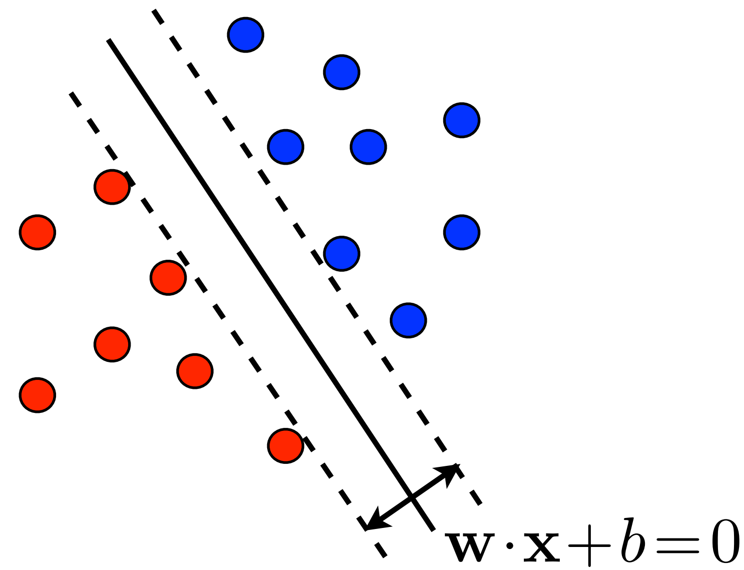
- Training data: sample drawn i.i.d. from set $X \subseteq \mathbb{R}^N$ according to some distribution D,

$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in X \times \{-1, +1\}.$$

- Problem: find hypothesis $h : X \mapsto \{-1, +1\}$ in H (classifier) with small generalization error $R_D(h)$.
- Linear classification:
  - Hypotheses based on hyperplanes.
  - Linear separation in high-dimensional space.

Linear Separation

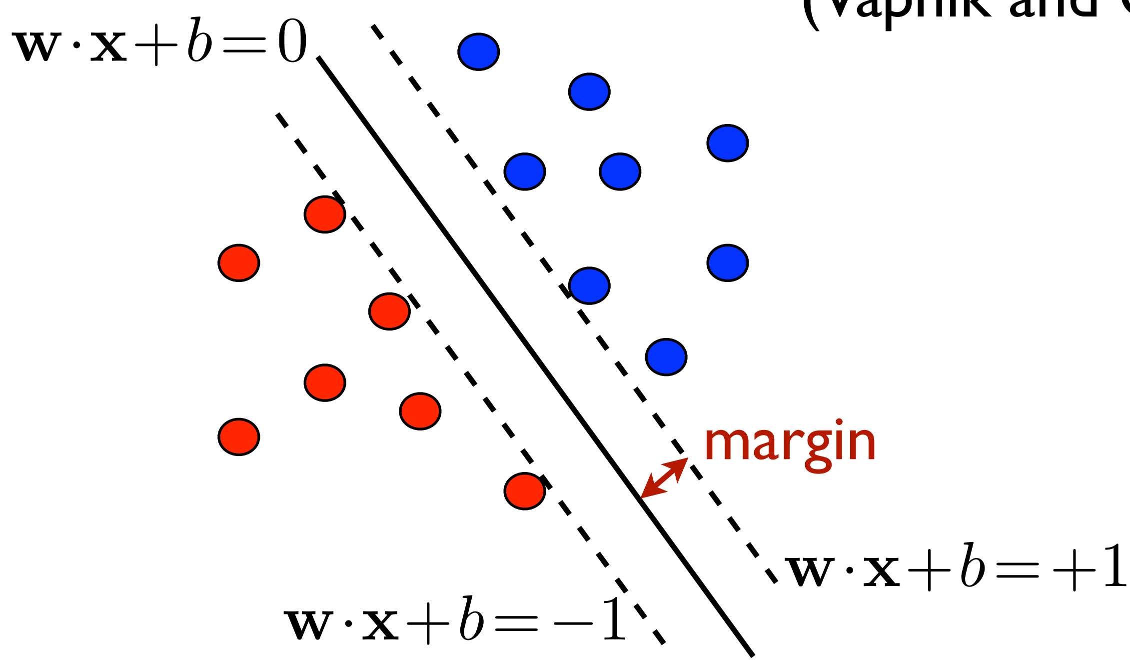




- Classifiers: $H=\{\mathbf{x}\mapsto \operatorname{sgn}(\mathbf{w}\cdot\mathbf{x}+b) : \mathbf{w}\in\mathbb{R}^N, b\in\mathbb{R}\}$.

Optimal Hyperplane: Max. Margin

(Vapnik and Chervonenkis, 1965)



- Canonical hyperplane: $\mathbf{w}$ and b chosen such that for closest points $|\mathbf{w}\cdot\mathbf{x}+b|=1$.
- Margin: $\rho=\min_{\mathbf{x}\in S}\frac{|\mathbf{w}\cdot\mathbf{x}+b|}{\|\mathbf{w}\|}=\frac{1}{\|\mathbf{w}\|}$.

Optimization Problem

- Constrained optimization:

$$\min_{\mathbf{w},b} \ \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to } \ y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1, i \in [1, m].$$

- Properties:
  - Convex optimization (strictly convex).
  - Unique solution for linearly separable sample.

Optimal Hyperplane Equations

- Lagrangian: for all $\mathbf{w}, b, \alpha_i \geq 0$,

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1].$$

- KKT conditions:

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \iff \boxed{\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i.}$$

$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \iff \boxed{\sum_{i=1}^{m} \alpha_i y_i = 0.}$$

$$\boxed{\forall i \in [1, m], \ \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0.}$$

Support Vectors

- Complementarity conditions:

$$\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i+b)-1]=0 \implies \alpha_i=0 \vee y_i(\mathbf{w}\cdot\mathbf{x}_i+b)=1.$$

- Support vectors: vectors $\mathbf{x}_i$ such that

$$\alpha_i \neq 0 \wedge y_i(\mathbf{w}\cdot\mathbf{x}_i+b)=1.$$

  - Note: support vectors are not unique.

Moving to The Dual

- Plugging in the expression of $\mathbf{w}$ in L gives:

$$L = \underbrace{\frac{1}{2}\|\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)}_{-\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)} - \underbrace{\sum_{i=1}^{m} \alpha_i y_i b}_{0} + \sum_{i=1}^{m} \alpha_i.$$

- Thus,

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j).$$

Dual Optimization Problem

- Constrained optimization:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subject to: } \alpha_i \geq 0 \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].$$

- Solution:

$$h(x) = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b \Big),$$

with $b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i)$ for any SV $\mathbf{x}_i$.

Leave-One-Out Analysis

- **Theorem**: let h_S be the optimal hyperplane for a sample S and let $N_{\mathrm{SV}}(S)$ be the number of support vectors defining h_S. Then,

$$\underset{S \sim D^m}{\mathrm{E}}[R(h_S)] \leq \underset{S \sim D^{m+1}}{\mathrm{E}}\left[\frac{N_{\mathrm{SV}}(S)}{m+1}\right].$$

- **Proof**: Let $S \sim D^{m+1}$ be a sample linearly separable and let $x \in S$. If $h_{S-\{x\}}$ misclassifies x, then x must be a SV for h_S. Thus,

$$\widehat{R}_{\mathrm{loo}}(\text{opt.-hyp.}) \leq \frac{N_{\mathrm{SV}}(S)}{m+1}.$$

Notes

- Bound on expectation of error only, not the probability of error.
- Argument based on sparsity (number of support vectors). We will see later other arguments in support of the optimal hyperplanes based on the concept of margin.

This Lecture

- Support Vector Machines - separable case
- Support Vector Machines - non-separable case
- Margin guarantees

Support Vector Machines

(Cortes and Vapnik, 1995)

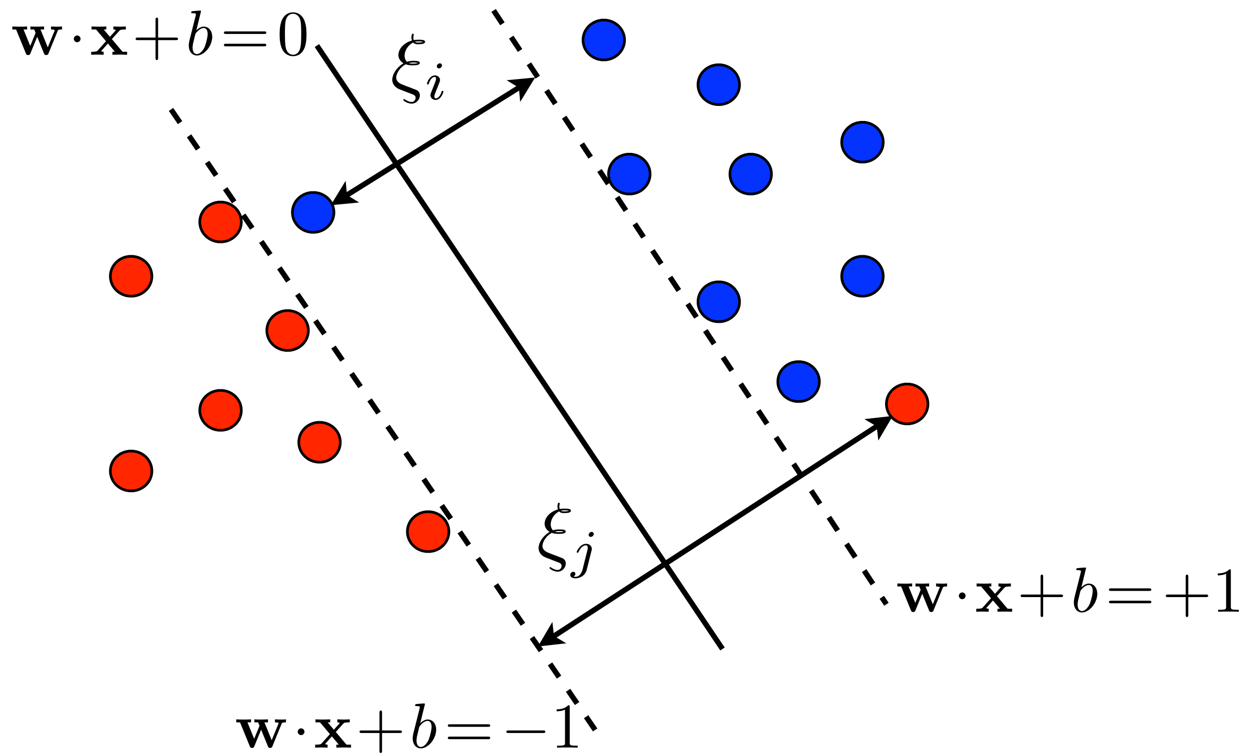
- Problem: data often not linearly separable in practice. For any hyperplane, there exists $\mathbf{x}_i$ such that

$$y_i\left[\mathbf{w}\cdot\mathbf{x}_i+b\right]\not\geq 1.$$

- Idea: relax constraints using slack variables $\xi_i \geq 0$

$$y_i\left[\mathbf{w}\cdot\mathbf{x}_i+b\right]\geq 1-\xi_i.$$

Soft-Margin Hyperplanes



- Support vectors: points along the margin or outliers.
- Soft margin: $\rho = 1/\|\mathbf{w}\|$.

Optimization Problem

(Cortes and Vapnik, 1995)

- Constrained optimization:

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

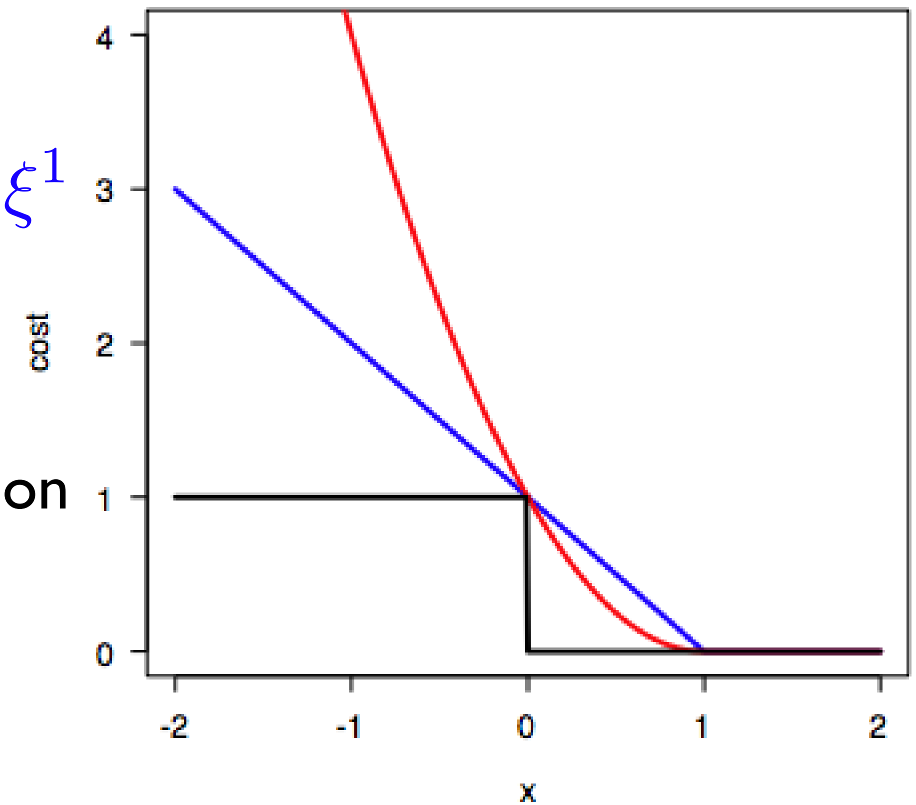
$$\text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i \;\wedge\; \xi_i \geq 0, i \in [1, m].$$

- Properties:
  - $C \geq 0$ trade-off parameter.
  - Convex optimization (strictly convex).
  - Unique solution.

Notes

- Parameter C: trade-off between maximizing margin and minimizing training error. How do we determine C?
- The general problem of determining a hyperplane minimizing the error on the training set is NP-complete (as a function of dimension).
- Other convex functions of the slack variables could be used: this choice and a similar one with squared slack variables lead to a convenient formulation and solution.

Hinge Loss



'Quadratic' hinge loss ξ^2
Hinge loss ξ^1
0/1 loss function
cost
4
3
2
1
0
-2
-1
0
1
2
x

SVMs Equations

- Lagrangian: for all $\mathbf{w}, b, \alpha_i \geq 0, \beta_i \geq 0,$

$$L(\mathbf{w}, b, \xi, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{m}\beta_i\xi_i.$$

- KKT conditions:

$$\nabla_w L = \mathbf{w} - \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i = 0 \iff \boxed{\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i.}$$

$$\nabla_b L = -\sum_{i=1}^{m}\alpha_i y_i = 0 \iff \boxed{\sum_{i=1}^{m}\alpha_i y_i = 0.}$$

$$\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \iff \boxed{\alpha_i + \beta_i = C.}$$

$$\boxed{\begin{array}{l} \forall i \in [1, m],\ \alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i] = 0 \\ \qquad\qquad \beta_i\xi_i = 0. \end{array}}$$

Support Vectors

- Complementarity conditions:

$$\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i] = 0 \implies \alpha_i = 0 \vee y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \xi_i.$$

- Support vectors: vectors $\mathbf{x}_i$ such that

$$\alpha_i \neq 0 \wedge y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \xi_i.$$

  - Note: support vectors are not unique.

Moving to The Dual

- Plugging in the expression of w in L gives:

$$L = \underbrace{\frac{1}{2}\|\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)}_{-\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)} - \underbrace{\sum_{i=1}^{m} \alpha_i y_i b}_{0} + \sum_{i=1}^{m} \alpha_i.$$

- Thus,

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j).$$

- The condition $\beta_i \geq 0$ is equivalent to $\alpha_i \leq C$.

Dual Optimization Problem

- Constrained optimization:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subject to: } 0 \leq \alpha_i \boxed{\leq C} \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].$$

- Solution:

$$h(x) = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b \Big),$$

with $b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i)$ for any $\mathbf{x}_i$ with $0 < \alpha_i \boxed{< C.}$

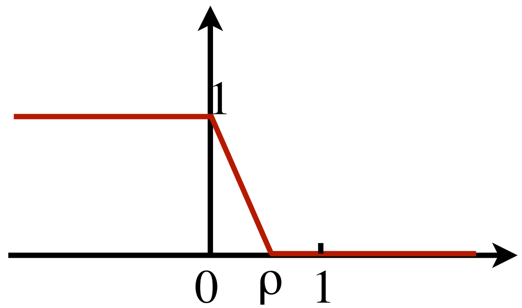
This Lecture

- Support Vector Machines - separable case
- Support Vector Machines - non-separable case
- Margin guarantees

Margin Loss

- Definition: for any $\rho > 0$, the ρ-margin loss is the function $L_\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ defined by for all $y, y' \in \mathbb{R}$ by $L_\rho(y, y') = \Phi_\rho(yy')$ with

$$\Phi_\rho(x) = \begin{cases} 0 & \text{if } \rho \leq x \\ 1 - x/\rho & \text{if } 0 \leq x \leq \rho \\ 1 & \text{if } x \leq 0. \end{cases}$$



- For a sample $S = (x_1, \ldots, x_m)$ and a hypothesis h, the empirical loss is

$$\widehat{R}_\rho(h) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(y_i h(x_i)) \leq \frac{1}{m} \sum_{i=1}^{m} 1_{y_i h(x_i) < \rho}.$$

Margin Bound - Linear Classifiers

- Corollary: Let $\rho > 0$ and $H = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} : \|\mathbf{w}\| \leq \Lambda\}$. Assume that $X \subseteq \{\mathbf{x} : \|\mathbf{x}\| \leq R\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}_\rho(h) + 2\sqrt{\frac{R^2 \Lambda^2 / \rho^2}{m}} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

High-Dimensional Feature Space

- Observations:
  - generalization bound does not depend on the dimension but on the margin.
  - this suggests seeking a large-margin separating hyperplane in a higher-dimensional feature space.
- Computational problems:
  - taking dot products in a high-dimensional feature space can be very costly.
  - solution based on kernels (next lecture).

References

- Corinna Cortes and Vladimir Vapnik, Support-Vector Networks, *Machine Learning*, 20, 1995.
- Koltchinskii, Vladimir and Panchenko, Dmitry. Empirical margin distributions and bounding the generalization error of combined classifiers. The Annals of Statistics, 30(1), 2002.
- Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer, New York.
- Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, Basederlin, 1982.
- Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.
- Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.